Deep Second-Order Siamese Network for Pedestrian Re-identification

Xuesong Deng[1,2], Bingpeng Ma[1,2], Hong Chang[1](✉), Shiguang Shan[1], and Xilin Chen[1]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
xuesong.deng@vipl.ict.ac.cn, bpma@ucas.ac.cn,
{changhong,sgshan,xlchen}@ict.ac.cn

[2] School of Computer and Control Engineering,
University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Typical pedestrian re-identification system consists of feature extraction and similarity learning modules. The learning methods involved in the two modules are usually designed separately, which makes them sub-optimal to each other, let alone to the re-identification target. In this paper, we propose a deep second-order siamese network for pedestrian re-identification which is composed of a deep convolutional neural network and a second-order similarity model. The deep convolutional network learns comprehensive features automatically from the data. The similarity model exploits second-order information, thus more suitable for re-identification setting than traditional metric learning methods. The two models are jointly trained over one unified large margin objective and the consistent convergence is guaranteed. Moreover, our deep model can be trained effectively with a small pedestrian re-identification dataset, through an irrelevant pre-training and relevant fine-tuning process. Experimental results on two public datasets illustrate the superior performance of our model over other state-of-the-art methods.

1 Introduction

Re-identifying a target pedestrian observed from a non-overlapping camera network is an important task in many real-world applications, such as threat detection, human retrieval and cross-camera tracking. During the past several years, it has drawn a lot of attentions from the field of computer vision and pattern recognition. Despite a lot of efforts spent on this task, pedestrian re-identification still remains largely unsolved due to low quality of images and complex variations in viewpoints, poses and illuminations. Some examples are shown in Fig. 1 to illustrate these difficulties.

Classical solutions for pedestrian re-identification mainly consist of two modules: feature extraction and similarity learning. Previous work usually focused on either feature extraction [1–10] or similarity learning [11–19].

S.-H. Lai et al. (Eds.): ACCV 2016, Part II, LNCS 10112, pp. 321–337, 2017.
DOI: 10.1007/978-3-319-54184-6_20

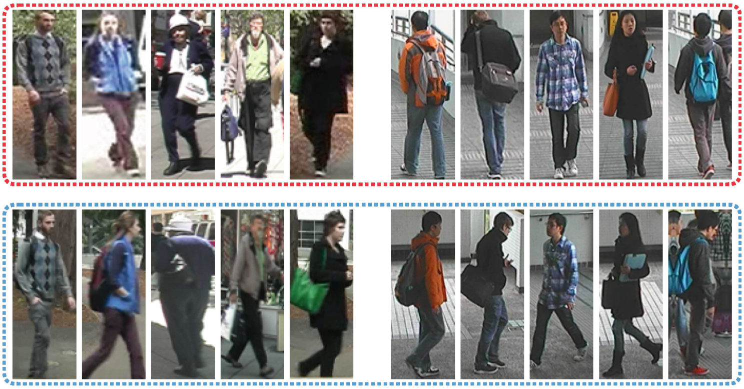


Fig. 1. Examples of pedestrian images. The left five columns are from VIPeR dataset [1], and the right ones are from CUHK01 dataset [16]. Each column contains two images from same person in different cameras, indicated by red and blue boxes respectively. (Color figure online)

Early feature extraction methods mainly capture two clues: color and texture. They combine sophisticated low-level features such as HSV, Gabor, HOG to describe the appearance model of pedestrian images, from which the similarity between a pair of images can then be measured by some similarity metric learning methods. Farenzena et al. [2] proposed the Symmetry-Driven Accumulation of Local Features (SDALF). They exploited the symmetry structure of pedestrians to handle view variations. Ma et al. [4] combined Gabor filters and covariance descriptor to handle illumination changes. However, handcrafted features are extremely hard to design due to the complicated variations in the real world camera networks. People often fail to take all interference factors into consideration. Recently, a series of work by Zhao et al. [7–9] attempted to use the saliency parts of pedestrians to estimate whether they are the same person or not. However, the saliency parts are not always consistent under different camera views. A common situation is that the same part of a pedestrian may look quite different from two cameras due to complex lighting conditions and pose changes. Recently Zheng et al. [20] proposed a general feature fusion scheme for image search which can also be utilized in pedestrian re-identification. Given a query image, their method can automatically evaluate the effectiveness of a to-be-fused feature, and then make use of the good features and ignore the bad ones.

Recently, some researchers try to boost the performance of pedestrian re-identification in the metric learning context [13–15,21]. Zheng et al. [13] proposed the Probabilistic Relative Distance Comparison (PRDC) model to maximize the likelihood of genuine pairs (from the same person) having smaller distances than those of imposter pairs (from different persons). A simple though effective strategy to learn a distance metric, named KISSME, from equivalence constraints was proposed in [14]. Mignon and Jurie [15] proposed the Pairwise Constrained Component Analysis (PCCA) to learn a projection from raw input space into a latent space where a desired constraint is satisfied. Chen et al. [21] introduced a mixture of linear similarity functions that is able to discover different matching patterns in the polynomial kernel feature map. These typical metric learning (ML) methods generally aim to automatically learn similarity metrics from data under supervised or semi-supervised learning setting. To this end, the original data is usually transformed to another feature space where the distance measure is more ideal for the learning objective. However, metric learning for pedestrian re-identification is much more difficult than that for traditional learning tasks such as classification. As for pedestrian re-identification problems, we need to estimate whether a pair of images are from the same person or not. For this purpose, metric learning here should deal with the challenges of large number of classes (persons) and large within-class variance. Moreover, since the training and testing datasets contain totally different persons, metric learning for pedestrian re-identification is expected to generalize well to unseen categories. Therefore, common metric learning strategy of globally linear transformation in Euclidean or Cosine distance may not be an effective solution for this problem. Different from previous typical methods, Li et al. [17] proposed the Locally Adaptive Decision Function (LADF) model that exploits second-order information which is more effective to model the complex relations between pedestrian image pairs. This model achieved good performance on pedestrian re-identification task.

It is worth noting that all the above methods only focused on one aspect of the pedestrian re-identification task, either feature extractor or similarity learning model. However, designing features and learning similarity model separately cannot guarantee their optimality for each other, thus making the whole re-identification system sub-optimal. If the feature extraction part fails to capture consistent and comprehensive features, even a sophisticated similarity learning algorithm will behave poorly. On the other hand, when we have features containing rich information, we still cannot re-identify a pedestrian successfully with an ineffective similarity learning model.

Recently, deep networks [22–24] are proposed to tackle this problem. The end to end architecture can improve the optimality of their models. Specifically, [22] proposed a siamese network named the Deep Metric which is quite similar to an early work by Chopra et al. [11]. The main difference is that they use the Cosine norm to evaluate the similarity score while Euclidean norm is used in [11]. This model is similar to typical metric learning except that Deep Metric takes raw image pixels (instead of handcrafted features) and transforms them nonlinearly through a convolutional neural network (CNN), instead of traditional linear transformations. Therefore, the CNN in Deep Metric model actually works as both a feature extractor and a metric learner. This is too much duty for a single model. Even with nonlinear approximation ability, Deep Metric model may not perfectly fit for pedestrian re-identification task. In Deep Re-ID [23] model, several layers are designed to tackle different issues in pedestrian re-identification, such as photometric transforms, displacement and pose transforms. However, it is too ideal to expect each single layer to handle a complex variation. Ahmed et al. [24] proposed an improved deep architecture for pedestrian re-identification. Similar to Deep Re-ID, they formulated the problem of pedestrian re-identification as a binary classification problem. And they both defined a similar layer to encode the differences between features, which are extracted from previous convolutional layers. As they both train their models directly with a relatively small pedestrian re-identification dataset, the scale of their networks is limited. This limitation results in weaker hierarchical semantic abstraction ability of their model.

To address all the above problems for pedestrian re-identification task, we propose a deep second-order siamese network (DSSN) which is composed of a CNN model as feature extractor and a second-order similarity model as similarity learner. Deep CNN model succeeds in many computer vision applications as a feature extractor thanks to its highly nonlinearity and hierarchical semantic abstraction ability. However, it is quite hard to train a large scale CNN model with a relatively small pedestrian re-identification dataset. Inspired by RCNN [25], we design an irrelevant pre-training and relevant fine-tuning strategy to initialize the deep CNN. In the similarity learning part, we propose a model which encodes second-order information. The higher-order similarity layer can model more complex relation than typical ML methods, thus more suitable for pedestrian re-identification task. Both the deep CNN and the second-order similarity model are trained alternately with one unified energy based loss function to guarantee their optimality for each other. Moreover the energy based loss function leads to a large margin solution, which in turn enhances the generalization ability of the proposed method. Experimental results on benchmark pedestrian re-identification tasks verify the effectiveness of our proposed method. The main contributions in this paper can be summarized as follows:

- With irrelevant pre-training and relevant fine-tuning process, we succeed in training a large scale deep CNN with a small pedestrian re-identification dataset.
- A similarity model encoding second-order information is proposed to estimate the similarities between feature pairs, which is more effective than previous metric learning methods for pedestrian re-identification problem.
- Thanks to the alternate learning strategy, a reasonably optimal large margin solution is guaranteed.

2 Deep Second-Order Siamese Network

To estimate the similarity between two pedestrian images, we propose a method named Deep Second-order Siamese Network (DSSN). The architecture of our proposed method is shown in Fig. 2. It is composed of two identical convolutional neural networks (CNN) and a second-order similarity (SS) function. With a carefully designed energy based loss function, an ideal large margin solution for re-identification task is guaranteed. In the following subsections, we present the CNN model, the second-order similarity model and the energy based large margin solution in details.

2.1 Convolutional Neural Network (CNN)

The re-identification performance is affected by many factors such as low image resolution, illumination, viewpoint and pose variations. To overcome these interference factors, we need to extract comprehensive and robust features. Designing features with all interference factors taken into consideration is extremely hard. Nevertheless deep CNN model shows its superior power over handcrafted models thanks to its high nonlinearity and hierarchical semantic abstraction in many computer vision applications. And with a elaborately designed objective, deep CNN model can be learned to fit for a certain goal. Thus we believe features learned from a deep CNN are more powerful than handcrafted features or features learned via shallow networks.

To this end, we construct a deep convolutional neural network as feature extractor for pedestrian re-identification task. The network is same as that proposed in [26] except that we remove the softmax layer. The deep CNN contains five convolutional layers and two fully connected layers. The forward process of each layer is expressed in the following equations:

$$\mathbf{z}^{l+1} = \mathbf{W}^{l+1} * \mathbf{a}^{l} + \mathbf{b}^{l+1} \tag{1}$$

$$\mathbf{a}^{l+1} = \sigma(\mathbf{z}^{l+1}) \tag{2}$$

$$\sigma(x) = \max(0, x) \tag{3}$$

$\mathbf{W}^{l+1}$ and $\mathbf{a}^{l+1}$ are the parameter matrix and the activation of the $(l+1)^{th}$ layer respectively. $\sigma(\cdot)$ is the activation function and we use Rectified Linear Units (ReLU) in this paper. The details about the network can be found in [26]. The outputs of the last fully connected layer are taken as the learned deep features.

As illustrated in Fig. 2, two pedestrian images are split into several non-overlapping stripes. The two CNNs take each pair of the corresponding stripes as input and output the learned deep features for the following similarity model.

2.2 Second-Order Similarity Function

In standard siamese network [11,22], the distance between two deep features is usually calculated by a simple Euclidean or Cosine metric. Previous experimental results [22] of this type of siamese network on pedestrian re-identification task did not demonstrate superior power over handcrafted feature followed by metric learning methods [17]. Actually, extracting general sophisticated features with a deep network is not good enough for pedestrian re-identification task. Proper metric learning with respect to high-level re-identification objective is still necessary.

To handle the complex relation addressed above, a variety of metric learning (ML) methods [11–19] have been proposed. Despite of different objective functions, these metric learning models calculate the new distance similarly as:

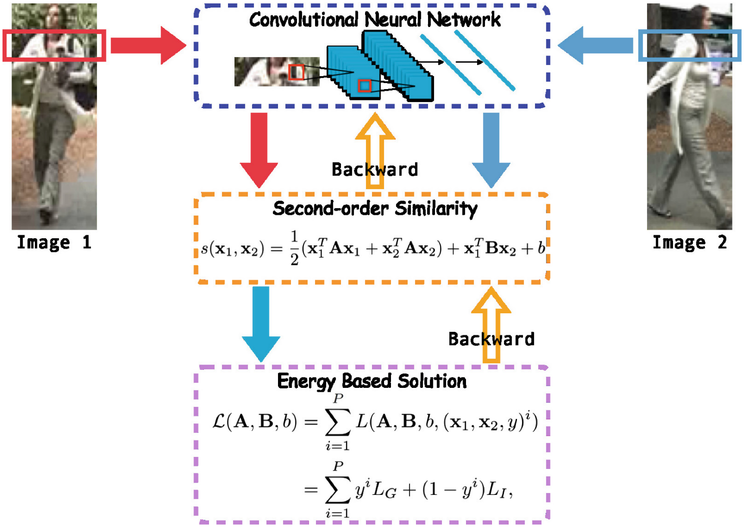

Convolutional Neural Network
Image 1
Image 2
Backward
Second-order Similarity
$s(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}(\mathbf{x}_1^T\mathbf{A}\mathbf{x}_1 + \mathbf{x}_2^T\mathbf{A}\mathbf{x}_2) + \mathbf{x}_1^T\mathbf{B}\mathbf{x}_2 + b$
Backward
Energy Based Solution
$\mathcal{L}(\mathbf{A}, \mathbf{B}, b) = \sum_{i=1}^{P} L(\mathbf{A}, \mathbf{B}, b, (\mathbf{x}_1, \mathbf{x}_2, y)^i) = \sum_{i=1}^{P} y^i L_G + (1 - y^i) L_I,$

Fig. 2. An overview of the deep second-order siamese network (DSSN). In the training phase, the gradients of the energy based loss function with respect to all model parameters are back-propagated through the whole network which are indicated by the yellow arrows. (Color figure online)

$$\begin{aligned} d_{ML}(\mathbf{x}_1, \mathbf{x}_2) &= (\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{L}^T \mathbf{L} (\mathbf{x}_1 - \mathbf{x}_2) \\ &= \mathbf{x}_1^T \mathbf{M} \mathbf{x}_1 + \mathbf{x}_2^T \mathbf{M} \mathbf{x}_2 - 2\mathbf{x}_1^T \mathbf{M} \mathbf{x}_2 \end{aligned} \tag{4}$$

$\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ are features extracted from a pair of pedestrian images. $\mathbf{M} = \mathbf{L}^T\mathbf{L}$ is a real symmetric matrix. By learning a desired $\mathbf{L}$, samples from the same class get closer to each other while those from different classes farther in a projected latent space.

However, as pointed out in [17], there is an intrinsic mismatch between typical ML methods and re-identification task. The projection matrix $\mathbf{L}$ learned from training samples may not work well for testing samples from new categories. A desired model for pedestrian re-identification requires the ability of adapting locally rather than a simple global projection. Similar to [17], we define a *second-order similarity function* to capture local data structures as follows:

$$s(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}(\mathbf{x}_1^T \mathbf{A} \mathbf{x}_1 + \mathbf{x}_2^T \mathbf{A} \mathbf{x}_2) + \mathbf{x}_1^T \mathbf{B} \mathbf{x}_2 + b. \tag{5}$$

$s(\mathbf{x}_1, \mathbf{x}_2) > 0$ means $\mathbf{x}_1$ and $\mathbf{x}_2$ being a genuine pair, otherwise an imposter pair. $\{\mathbf{A}, \mathbf{B}, b\}$ are parameters for the similarity function. $\mathbf{A}$ and $\mathbf{B}$ are real symmetric matrices, and b is the bias term. And due to the symmetric constraint: $s(\mathbf{x}_1, \mathbf{x}_2) = s(\mathbf{x}_2, \mathbf{x}_1)$, we use $\mathbf{A}$ for both $\mathbf{x}_1$ and $\mathbf{x}_2$. Compared with typical ML methods, the second-order similarity metric can model much more complex relations due to three sets of parameters $\{\mathbf{A}, \mathbf{B}, b\}$.

2.3 Energy Based Large Margin Model

To learn improved deep features and an ideal second-order similarity function for pedestrian re-identification task, we adopt an energy based large margin model as it can generalize well to unseen examples. Suppose there are P pairs of labeled images, $(\mathbf{x}_1, \mathbf{x}_2, y)^i, i = 1, \ldots, P$, where $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ are the features extracted by the CNN model described before, and $y^i = 1$ indicates a genuine pair and $y^i = 0$ an imposter pair. We define the overall loss function as:

$$\begin{aligned}\mathcal{L}(\mathbf{\Theta}) &= \sum_i^P L(\mathbf{\Theta}, (\mathbf{x}_1, \mathbf{x}_2, y)^i) \\ &= \sum_i^P y^i L_G(\mathbf{\Theta}, E(\mathbf{x}_1^i, \mathbf{x}_2^i)) + (1 - y^i) L_I(\mathbf{\Theta}, E(\mathbf{x}_1^i, \mathbf{x}_2^i))\end{aligned} \tag{6}$$

$\mathbf{\Theta}$ represents the whole set of parameters involved in the CNN and the similarity function. $L_G(\cdot)$ and $L_I(\cdot)$ are the loss functions for genuine and imposter pairs respectively. $E(\mathbf{x}_1^i, \mathbf{x}_2^i)$ is the energy function measuring the compatibility of a pair of the image features, which is defined as:

$$E(\mathbf{x}_1^i, \mathbf{x}_2^i) = \frac{1}{Z} \exp\left(-\frac{s(\mathbf{x}_1^i, \mathbf{x}_2^i)}{\lambda_0}\right), \tag{7}$$

where

$$Z = \sum_{i=1}^P \exp\left(-\frac{s(\mathbf{x}_1^i, \mathbf{x}_2^i)}{\lambda_0}\right) \tag{8}$$

is the partition function. Lower energy $E(\mathbf{x}_1^i, \mathbf{x}_2^i)$ indicates larger $s(\mathbf{x}_1^i, \mathbf{x}_2^i)$, which suggests $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ form a genuine pair. On the contrary, higher energy suggests the data being an imposter pair.

The two partial loss functions $L_G(\cdot)$ and $L_I(\cdot)$ take the forms of:

$$L_G(\mathbf{x}_1^i, \mathbf{x}_2^i) = \alpha E(\mathbf{x}_1^i, \mathbf{x}_2^i)^2, \tag{9}$$

$$L_I(\mathbf{x}_1^i, \mathbf{x}_2^i) = 2\exp\left(-\beta E(\mathbf{x}_1^i, \mathbf{x}_2^i)\right) \tag{10}$$

α and β are two constant parameters. The energy value $E > 0$ is guaranteed due to its definition. Clearly, $L_G(\cdot)$ is a monotonically increasing function and $L_I(\cdot)$ is a monotonically decreasing function, respect to E. It has been proved in [11] that with the partial losses of such monotonicity, minimizing the loss function $\mathcal{L}(\mathbf{\Theta})$ defined in Eq. (6) leads to a large margin solution which can generalize well to unseen persons.

To understand in a more loose way, minimizing $\mathcal{L}(\mathbf{\Theta})$ corresponds to decreasing the energy $E(\mathbf{x}_1^i, \mathbf{x}_2^i)$ for genuine pairs and increasing the energy for imposter pairs. Equivalently, the objective leads to larger similarity score $s(\mathbf{x}_1^i, \mathbf{x}_2^i)$ for genuine pairs and smaller similarity score for imposter pairs.

Researchers in [17] also proposed a method to approach a large margin solution for second-order metric model. They train their proposed model in SVM-like fashion. It should be noticed that the SVM-like solution is not quite straightforward to combine with the back-propagation algorithm. That's why we choose the energy-based loss function.

2.4 Gradients

The parameters in both CNN and second-order similarity model are jointly optimized with the unified loss function 6. We use alternate training process to adjust each part of the deep second-order siamese network to get an optimal model. It should be pointed out that Li et al. [17] proposed a SVM-like objective function to approach the large margin solution. However, it is pretty hard for the SVM-like objective to optimize the CNN model. Instead, we adopt the energy based loss function, since it is quite straightforward to use back propagation (BP) algorithm to optimize both the CNN and similarity models under the energy based framework.

Optimizing the second-order similarity model is pretty straightforward. We can calculate the derivatives of the loss function (6) with respect to $\{\mathbf{A}, \mathbf{B}, b\}$ as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = \sum_{i=1}^{P} \nabla^i * \frac{\mathbf{x}_1^i {\mathbf{x}_1^i}^T + \mathbf{x}_2^i {\mathbf{x}_2^i}^T}{2} \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{B}} = \sum_{i=1}^{P} \nabla^i * \mathbf{x}_1^i {\mathbf{x}_2^i}^T \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{P} \nabla^i \tag{13}$$

with

$$\nabla^i = \frac{2y^i \alpha (E^3 - E^2) + 2(y^i - 1)\beta (E^2 - E) \exp(-\beta E)}{\lambda_0}, \tag{14}$$

where E is the abbreviation for $E(\mathbf{x}_1^i, \mathbf{x}_2^i)$. Then, we can use a gradient based optimization method like L-BFGS to obtain local optimal estimates for $\mathbf{A}, \mathbf{B}$ and b.

We use back-propagation (BP) algorithm to optimize the parameters of the convolutional neural network. Suppose the CNN has N layers and $\mathbf{W}^n$ is the parameter matrix of the n-th layer. The gradients of loss $\mathcal{L}$ with respect to $\mathbf{W}^n$ is calculated as:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^n} = \mathbf{a}^{n-1} \delta^n \tag{15}$$

$$\delta^N = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^N} \odot \sigma'(\mathbf{z}^N) \tag{16}$$

$$\delta^n = ({\mathbf{W}^{n+1}}^T \delta^{n+1}) \odot \sigma'(\mathbf{z}^n) \tag{17}$$

$\mathbf{z}^n$ is the input of the activation function $\sigma(\cdot)$ in the n-th layer. $\mathbf{a}^n$ is the corresponding activation vector. δ^n is the *error term* defined in BP algorithm.

The inputs to the second-order metric model, $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ are also the outputs of the CNNs. Thus δ^N in our model can be calculated as:

$$\delta^N = \sum_i \frac{\partial \mathcal{L}}{\partial \mathbf{x}^i} \odot \sigma'(\mathbf{z}^N) \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_1^i} = \nabla^i * (\mathbf{A}\mathbf{x}_1^i + \mathbf{B}\mathbf{x}_2^i) \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_2^i} = \nabla^i * (\mathbf{A}\mathbf{x}_2^i + \mathbf{B}\mathbf{x}_1^i) \tag{20}$$

Once we get δ^N, the rest work can be done with typical BP algorithm.

As proved in [11], minimizing the loss function (6) with respect to the parameters of the CNN and the second-order similarity model will both lead to large margin solutions. Therefore, consistent convergence of our optimization method is guaranteed.

3 Learning Strategy

As large scale of parameters generally requires large scale of training samples, successful deep networks are usually trained on large datasets, such as ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [27] which contains millions of images. However, most of the public pedestrian re-identification datasets only contain several thousand images. It is unreasonable to directly train a deep model involving millions of parameters with such a small dataset.

In the work of Girshick et al. [25], researchers found a way to bridge the huge gap between small scale dataset and large scale model. They pre-train the model on ILSVRC dataset with supervision and then fine-tune the model with a domain-specific loss function on a small dataset. The reason behind the success is that lower level convolutional filters detecting edges, orientations, *etc.* can be shared across different sources of datasets, while higher level filters encoding rich semantic information must be task specific. Pre-training on a large dataset with supervision can generate effective low-level filters and a good initialization for high-level filters. After that fine-tuning process can further adjust the model to fit some specific task on a small dataset.

Inspired by this paradigm, our learning strategy consists of three stages. First, we pre-train the CNN model on the ILSVRC dataset. On the second stage, we fine-tune it with a pedestrian dataset with supervision. At last, we jointly minimize the loss function (6) with respect to both the CNN model and the second-order similarity model.

3.1 Supervised Pre-training and Fine-Tuning CNN

We pre-train the CNN model with the open source Caffe library [28] on ILSVRC dataset. The detailed training process is the same as [25]. After the pre-training

Algorithm 1. The main learning algorithm

```
Input: Dataset $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{x}_2, y)^i\}_{i=1}^n$
    Random initialized $\mathbf{W}$ (CNN model)
    Random initialized $\{\mathbf{A}, \mathbf{B}, b\}$ (SS model)
Output: $\mathbf{W}_{opt}$ and $\{\mathbf{A}, \mathbf{B}, b\}_{opt}$
Pre-train $\mathbf{W}$ on ILSVRC dataset
Fine-tune $\mathbf{W}$ on $\mathcal{D}$ with Euclidean norm as similarity measure
while until convergence do
    Randomly select a batch of data samples from $\mathcal{D}$
    Compute the value of Eq. 6 for this batch
    Use back-propagation algorithm to calculate the gradients of $\mathbf{W}$ and $\{\mathbf{A}, \mathbf{B}, b\}$
    Update $\mathbf{W}$ and $\{\mathbf{A}, \mathbf{B}, b\}$
end while
$\mathbf{W}_{opt} \leftarrow \mathbf{W}$
$\{\mathbf{A}, \mathbf{B}, b\}_{opt} \leftarrow \{\mathbf{A}, \mathbf{B}, b\}$
return $\mathbf{W}_{opt}$ and $\{\mathbf{A}, \mathbf{B}, b\}_{opt}$
```

step, the lower level convolutional filters are capable of extracting general instructive low-level image features.

The high-level filters in the pre-trained CNN model is suitable for image classification but may not fit for encoding significant information for pedestrian re-identification. Therefore, we fine-tune the CNN model using a set of pedestrian images. To this end, we construct genuine pairs and imposter pairs from these images, and minimize the loss function $\mathcal{L}(\mathbf{\Theta})$ on these sample pairs. In this stage, we replace the second-order similarity model in the DSSN model with a Euclidean norm. The effectiveness of the learned CNN model for pedestrian re-identification task is validated through experiments in Sect. 4.3.

3.2 Joint Optimization

After the fine-tuning process, we need to minimize the loss function (6) with respect to both the CNN and the second-order similarity model to guarantee their optimality for each other. To achieve this goal, we calculate the gradients of the second-order similarity model based on Eqs. 11–13. And the gradients of CNN in each layer can be computed through back-propagation algorithm. The error terms in the last full connected layer is in the form of Eqs. 18–20. Classical stochastic gradient descent (SGD) algorithm can be applied to optimize the CNN and second-order similarity model simultaneously. As discussed in Sect. 2.3, each step in the optimization process will lead to a large margin solution. Thus the DSSN model composed of the CNN model and second-order similarity model is guaranteed to converge to a local optimal state, which is fairly desirable for the pedestrian re-identification task.

To summarize, our overall learning algorithm is described in Algorithm 1.

4 Experiments and Analysis

We conduct our experiments on two benchmark datasets, *i.e.* the VIPeR [1] dataset and the CUHK01 [16] dataset. We compare our method to other deep models and state-of-the-art methods for pedestrian re-identification. The results validate the effectiveness of our proposed DSSN model.

4.1 Datasets

VIPeR dataset contains 632 people and 2 images for each pedestrian. Images are normalized to 48×128 for evaluations. A pair of images are captured from 2 different camera views (camera A and camera B). The viewpoint change is of 90° or more. Complex illumination conditions and huge pose variations make VIPeR dataset the most challenging pedestrian re-identification dataset.

CUHK01 dataset contains 971 people which are also captured from 2 camera views and are normalized to 60×160. And there are 2 images for each pedestrian in each camera view. Images in CUHK01 dataset have higher resolution. The illumination condition is more stable than VIPeR dataset. The better quality makes it possible for the DSSN to encode more information.

4.2 Evaluation Protocol

Our experiments on both datasets follow the evaluation protocol in [1]. The datasets are randomly partitioned into two even parts as training set and testing set. For VIPeR dataset 316 pedestrians are randomly picked up as training samples and 486 pedestrians for CUHK01 dataset. In the testing phase, the probe set is composed of images from camera A, and the gallery set is from camera B. We calculate the similarity scores between a target in the probe set and all candidates in the gallery set based on our proposed model. Then we can get the ranking of the candidates based on their similarity scores. The standard cumulative matching characteristic (CMC) curve is then reported to measure the performance over the whole probe set [29]. Generally higher CMC curve indicates better performance. To get stable statistics, all experiments are repeated 10 times with random training and testing partition on both datasets. And the average CMC curves over 10 trials are reported to evaluate the performance on both datasets.

4.3 Feature Learned via Deep CNN V.S. Handcrafted Feature

To evaluate the effectiveness of the deep CNN model, we compare the feature extracted from the fine-tuned deep CNN model (*Deep* feature) with several handcrafted features. They are HGR feature [30], eLDFV [3], eBiCov [4] and QALF [20]. Among them, HGR feature is used in [17] and achieved a good performance on pedestrian re-identification task. It is a hierarchical gaussianization representation based on simple patch color descriptors. It can be seen as the baseline of the handcrafted feature. eLDFV is a fusion of Weighted Color Histograms (wHSV), Maximally Stable Color Regions (MSCR) [2] and fisher vectors encoded local descriptors. eBiCov is a bio-inspired covariance descriptor fused with wHSV and MSCR. QALF is a self-adaptive feature fusion method that fuses several low-level features such as Color Histograms, Color Names, LBP, HOG. The results of different feature methods on VIPeR dataset are presented in Table 1. The best results at each rank are highlighted in bold face.

Table 1. Comparison with different features on VIPeR (Unit: %).

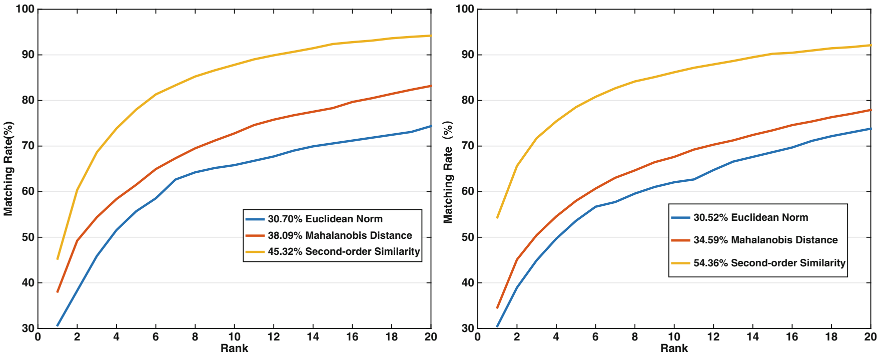
Method	Rank-1	Rank-5	Rank-10	Rank-20
HGR	9.46	22.50	31.80	40.41
eLDFV	22.34	46.92	60.04	71.81
eBiCov	20.66	42.62	56.11	67.67
QALF	30.17	51.60	62.44	73.81
Deep feature	**30.70**	**55.70**	**65.82**	**74.37**

From Table 1, we can tell that the *Deep* feature is much better than HGR, eLDFV and eBiCov features. The Rank-1 matching rate is around 20%, 8% and 10% higher than these three methods respectively. It is worth noting that eLDFV and eBiCov both fused several different features. And fusing features together generally can achieve better performance. QALF can even reward good features with higher weights and punish bad features. But the *Deep* feature alone is still slightly better than the QALF feature. These results demonstrate that due to the great semantic abstraction ability of the deep CNN model, the learned *Deep* feature is better than handcrafted ones for pedestrian re-identification task.

4.4 Comparison Between Similarity Methods

In our proposed model, the second-order similarity (SS) function is defined to model the complex relation between pair data. To validate the need for metric learning and the effectiveness of our SS model, we compare it with other similarity methods including Euclidean norm (Euc) and Mahanalobis based distance function (Eq. 4). Euc is compared as the baseline method. And Mahalanobis based distance is the similarity measure in most typical metric learning methods. We use Euc and Mahalanobis based distance as similarity function respectively to replace our SS model. And they are jointly trained with the CNN model under the same strategy described in Sect. 3. The results of different similarity methods on VIPeR and CUHK01 datasets are shown in Fig. 3.

It is clear showed in Fig. 3(a) and (b) that the performance of Mahalanobis based distance are better than Euc. The Rank-1 matching rate of Mahalanobis based distance are around 8% and 4% higher than Euc on VIPeR and CUHK01 respectively. While adopting the same CNN model and training strategy, the



(a) CMC curves on VIPeR (b) CMC curves on CUHK01

Fig. 3. Experimental results compared with different similarity methods.

only difference between these methods are the similarity measures. The better performance demonstrate that even processed by a strong CNN model, metric learning method is still a essential part in pedestrian re-identification.

We also notice that our proposed SS model achieves much better results than Mahalanobis based distance. The Rank-1 matching rate on two datasets are around 7% and 20% higher respectively. Mahalanobis based distance, like other typical ML method tries to find an ideal latent space through one single global projection while our SS model exploits the second-order information. Hence the better performance demonstrates the advantage of the second-order model over typical ML methods. This is all due to the ability of modeling more complex relation.

4.5 Comparison with Other Deep Models

As the proposed DSSN model is a deep model, we compare it with other deep models proposed for pedestrian re-identification task. These models include Deep Metric [22], FPNN [23] and Improved Deep [24]. Deep Metric is a siamese network but with similarity layer being a simple Cosine norm. FPNN model and Improved Deep model both formulate pedestrian re-identification task as a binary classification problem and train their models directly with relative small pedestrian re-identification datasets.

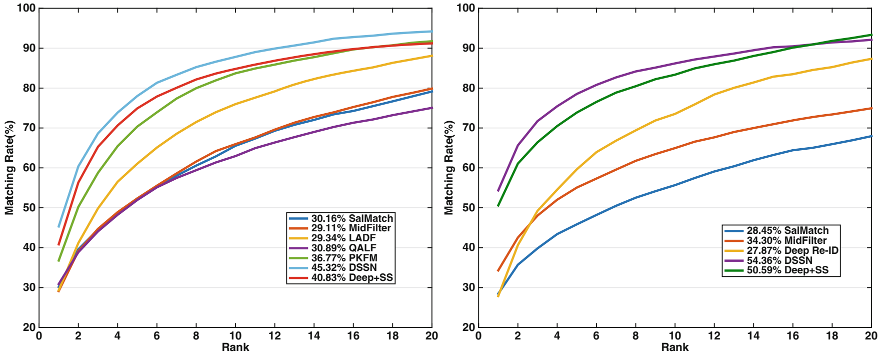
In Table 2, we present the experimental results of different deep models on VIPeR and CUHK01 datasets. Unavailable statistic data is denoted by −. And the best results at each rank are highlighted in bold face. From Table 2 it is obvious that our proposed DSSN achieves the best performance on both datasets. On VIPeR dataset, the Rank-1 matching rate of our DSSN model is 45.31% while the best of other methods is only 34.81%. This shows the superior power of our DSSN model over other deep models. Besides we notice that our DSSN model gets more improvement on VIPeR dataset than on CUHK01 dataset. On VIPeR dataset the Rank-1 matching rate is around 11% higher than the others while

Table 2. Comparison with deep models on VIPeR and CUHK01 (Unit: %).

Method	VIPeR				CUHK01			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
Deep metric	28.23	59.27	73.45	86.39	-	-	-	-
Improved deep	34.81	63.61	75.63	84.49	47.53	71.60	80.25	87.45
FPNN	-	-	-	-	27.87	59.64	73.53	87.34
DSSN	**45.31**	**78.00**	**87.81**	**92.37**	**54.36**	**78.53**	**86.19**	**92.11**

it is only around 7% on CUHK01 dataset. Considering the fact that VIPeR dataset suffers more from viewpoint changes and illumination variations than CUHK01 dataset, the larger improvement on VIPeR dataset indicates that our DSSN model is more robust to these interference factors than the other ones.

Compared with the Cosine norm of Deep Metric, our second-order similarity model can model more complex relation. And as for FPNN and Improved Deep they train their network directly while our DSSN adopts the pre-training and fine-tuning process. This makes our model deeper and better trained than theirs. Further more, the joint optimization leads our model to a overall optimal state. These differences result in the remarkable improvement over other deep models.



(a) CMC curves on VIPeR

(b) CMC curves on CUHK01

Fig. 4. Experimental results compared with state-of-the-art methods.

4.6 Comparison with State-of-the-Art Methods

After validating the effectiveness of each part in our DSSN model, we also compare our DSSN model with several state-of-the-art methods: SalMatch [8], MidFilter [9], LADF [17], QALF [20] and PKFM [21]. SalMatch and MidFilter define features based on the saliency parts in pedestrian images. LADF solved a SVM-like objective which leads its similarity model to a large margin solution. PKFM introduced a mixture of linear similarity functions to discover different matching patterns in the polynomial kernel feature maps.

CMC curves on VIPeR and CUHK01 dataset of different methods are shown in Fig. 4(a) and (b) respectively. We denote the methods of combining the *Deep* feature with our SS model as Deep+SS. The difference between Deep+SS and our full DSSN model is that the CNN model and the SS model of Deep+SS method are trained separately while our DSSN model are jointly optimized. From Fig. 4(a) we can see that on VIPeR dataset Deep+SS method achieves better result than other methods except for DSSN. The Rank-1 matching rate of PKFM is only 36.77% while Deep+SS method reaches at 40.83%. This result indicates that the fine-tuned CNN model and the SS model are quite effective. Simple combination achieves decent improvement over other methods. Furthermore, the performance of DSSN are better than Deep+SS on a remarkable scale. The Rank-1 matching rate of DSSN reaches at 45.31%. This indicates the proposed joint optimization algorithm further improves the CNN model and the second-order similarity model.

Similar results have been found on CUHK01 dataset in Fig. 4(b). Deep+SS outperforms others while DSSN achieves the best result. The Rank-1 matching rate of DSSN is 54.35%. These results further validate the effectiveness of the proposed DSSN model and the joint optimization algorithm.

5 Conclusion

In this paper, we propose a novel deep second-order siamese network for pedestrian re-identification which consists of feature extraction and similarity learning modules. The features learned via the deep CNN model encode effective information for pedestrian re-identification. And the second-order relation exploited in the similarity function makes the model more suitable for re-identification task. We propose an joint optimization process to train the model successfully. Therefore the feature learning and similarity learning modules are optimal for each other, which is rarely seen in previous related works. The experimental results validate the superiority of our model over other methods on two benchmark datasets.

Acknowledgment. This work was supported in part by National Basic Research Program of China (973 Program): 2015CB351802, and Natural Science Foundation of China (NSFC): 61272319, 61390515 and 61572465.

References

1. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_21

2. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. IEEE (2010)
3. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 413–422. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_41
4. Ma, B., Su, Y., Jurie, F.: BiCov: a novel image representation for person re-identification and face verification. In: British Machine Vision Conference, 11 p. (2012)
5. Li, W., Wang, X.: Locally aligned feature transforms across views. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3594–3601. IEEE (2013)
6. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1622–1634 (2013)
7. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3586–3593. IEEE (2013)
8. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2528–2535. IEEE (2013)
9. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 144–151. IEEE (2014)
10. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 536–551. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_35
11. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 539–546. IEEE (2005)
12. Prosser, B., Zheng, W.S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: BMVC, vol. 2, p. 6 (2010)
13. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 649–656. IEEE (2011)
14. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2288–2295. IEEE (2012)
15. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2666–2672. IEEE (2012)
16. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37331-2_3
17. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3610–3617. IEEE (2013)

18. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 653–668 (2013)
19. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_1
20. Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
21. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1565–1573 (2015)
22. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 34–39. IEEE (2014)
23. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 152–159. IEEE (2014)
24. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. Differences **5**, 25 (2015)
25. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587. IEEE (2014)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
27. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
28. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
29. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3. Citeseer (2007)
30. Zhou, X., Cui, N., Li, Z., Liang, F., Huang, T.S.: Hierarchical gaussianization for image classification. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1971–1977. IEEE (2009)